DOCUMENT RESUME

ED 143 373 IR 005 177

AUTHOR Deutsch, Donald R.; Kraft, Donald H.
TITLE A Study of an Information Retrieval Performance Measure: Expected Search Length as a Function of File Size and Organization.
PUB DATE 74
NOTE 27p.; Paper presented at the annual meeting of the Operations Society of America (Boston, Massachusetts, 1974)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Evaluation; *Indexing; *Information Retrieval; Information Storage; *Information Systems; *Information Theory; Libraries; Literature Reviews; Measurement; *Models; Organization

ABSTRACT

A framework is developed to evaluate expected search length, an important measure of an information storage and retrieval system's performance, as a function of the size of the information file and its organization in terms of indexing search structure. Previous research pertaining to search length by W.S. Cooper and others is surveyed and extended. A conceptualization recognizing a dichotomy of retrieval activities--intellectual access versus physical access--is presented. Three cases are analyzed to demonstrate the benefits of indexing and organizing an information file. The applicability of such a model to library systems is demonstrated and potential areas for future research are identified. (Author/KP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED143373

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Donald H. Kraft

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND THE ERIC SYSTEM CONTRACTORS."

A Study of An Information Retrieval Performance Measure: Expected Search Length as a Function of File Size and Organization

Donald R. Deutsch and Donald H. Kraft

IR005177

A Study of An Information Retrieval Performance Measure: Expected Search Length as a Function of File Size and Organization

Donald R. Deutsch* and Donald H. Kraft**

University of Maryland
College Park, Maryland

Abstract

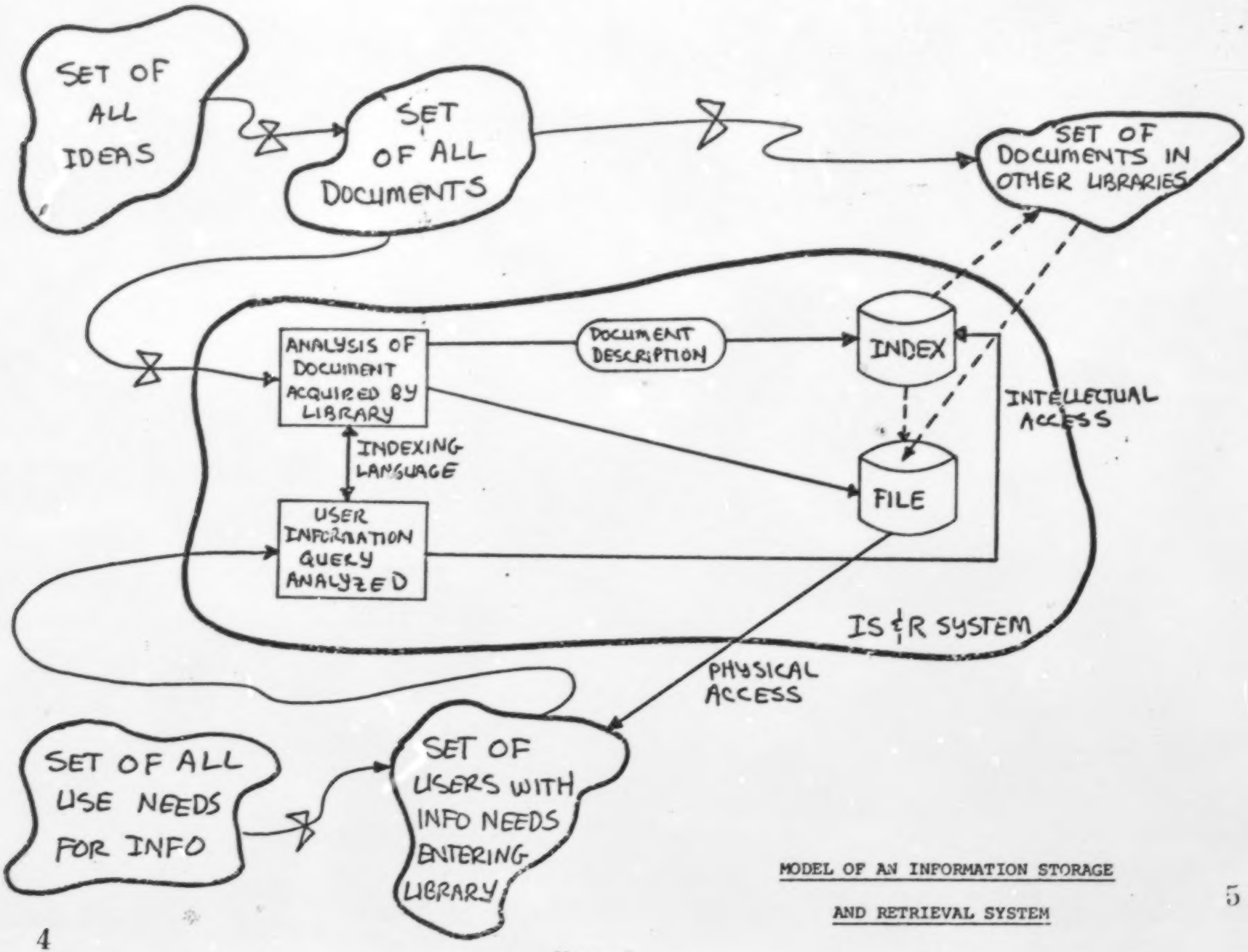
A framework is developed to evaluate expected search length, an important measure of an information storage and retrieval system's performance, as a function of the size of the information file and its organization in terms of indexing search structure. Previous research pertaining to search length by Cooper and others is surveyed and extended. A conceptualization recognizing a dichotomy of retrieval activities is presented. Three cases are analyzed to demonstrate the benefits of indexing and organizing an information file. Moreover, the applicability of such a model to library systems is demonstrated and potential areas for future research are identified.

Introduction

Information storage and retrieval (I.S.& R.) systems have been widely discussed in the literature. The nature and structure of I.S.& R. systems are discussed by Becker and Hayes (3), and are illustrated in Figure 1. A subset of all information is identified, selected, and acquired into the the system. Attributes or characteristics of each acquired information item are analyzed and recorded in an index, while the information itself is addressed and stored in a file for future retrieval. Utilization occurs when

* Department of Information Systems Management

** College of Library and Information Services



MODEL OF AN INFORMATION STORAGE AND RETRIEVAL SYSTEM

figure 1

from the set of all users requiring information, a subset of clients with specific needs enters the system. User queries are analyzed and matched against the index to determine a set of relevant information items which should satisfy the users' requirements. The user is then provided with the list of items deemed relevant to his query, abstracts of the selected information, or perhaps even the information items themselves. In any case, the cited information must be located and taken from storage.

Note that this characterization of I.S.& R. systems is consistant with, but is not limited in its application to, a traditional library with a collection of books and journals. Rather than considering documents, we have described the acquisition, storage, and retrieval of information, a more general term. Similarly, we have not specified the mechanisms used for index maintenance and reference or for storing the information itself; the automation of one or both of these system components is neither required nor precluded.

A major intellectual problem is the design of an indexing language (a vocabulary of acceptable terms and a syntax or set of grammatical rules for using the vocabulary) that allows communication between the searcher and the indexing mechanism in determining which information items are relevant to a specific user request. Another problem is the definition of what is relevant or pertinent.

One of the difficulties facing the systems analyst and designer is how to evaluate the performance of a specific retrieval system configuration. This is vital for comparison of the desirability of various alternative systems, and for evaluating existing system performance. The literature (9,13,16) contains many examples of criteria for library system effectiveness ratios, capability, beneficial effects, quality, value, and search length. The problem is

to adequately define and to operationalize each criterion so that it can be measured. Often, consideration is given only to one criterion, perhaps related to only one aspect of retrieval, rather than to an overall view. Moreover, the data used to evaluate the system may not fully reflect the criterion it is supposed to measure. Papier (17) suggests that there are six sources of criteria which should be examined (examples of each appear in parentheses): internal environment (formal and informal organization), external environment (technology, competition), user environment (system output), content area (quality, special requirements), system area (cost, time, volume) and vitality area (innovation, creativity, survival).

Cooper (8) user satisfaction is the "ideal" criterion, measured by the users' subjectively estimated utility for each document retrieved in a search sequence. He argues persuasively that there is considerable justification for the criterion of subjective user evaluation, but that its scope is limited solely to overall retrieval benefits; moreover, it is limited to existing systems, requires many large samples of users, and is both delicate and time consuming. Cooper recommends validation by experiment of any approximation measure, and insists on isolating the assumptions underlying any such measure of retrieval systems performance. He suggests search length as a practical measure that approximates the ideal and points out the assumptions required in order to use this measure as a surrogate for user satisfaction.

The purposes of this paper are to review the work of Cooper and others pertaining to search length as a measure of I.S.& R. system performance, to extend this concept so that different classes of search effort can be recognized when evaluating existing and proposed library systems, and to identify potential areas for future research.

A Dichotomy: Intellectual versus Physical Access

The retrieval process can be viewed as consisting of two separate steps. First, the user must obtain intellectual access to the system. That is, knowledge of the existence of information relevant to a particular query and of the locations or identifying labels of pertinent items is required. Intellectual access is provided by the indexing mechanism.

This mechanism may be a simple manual card or book catalog, or a sophisticated computerized reference system; the purpose is the same regardless of the complexity of the system used. There are several levels of complexity for I.S.& R. system automation; one can merely computerize the steps required when referencing a system manually, or one can go far beyond manual procedures. In systems with a moderate level of automation, descriptors from a limited vocabulary are combined with the Boolean operators "and", "or", and "not" to form a description of the desired information items.

Once the user is aware of the existence and location of relevant information items, he must obtain physical access to the information. Physical access involves the actual movement of the information from its location in storage to the user. Again, this may be either a strictly manual operation or an automated one. In a traditional library, physical access requires that the library staff or user locate and remove the described item from the storage stocks. In a system where the information is maintained in machine readable form, physical access involves the location, retrieval from on-line storage, and output of the information in a form suitable for human use.

Complex automated systems must actually locate the pertinent information items stored on external on-line storage devices. Various mechanisms are used

depending on the physical storage structure. For example, a direct relation storage structure is one where the indentifying key indicates the address of the storage media while an index sequential storage structure requires first a table search to determine the segment of storage, and then a scan of the contents of that segment to find the pertinent item. Other methods of converting identifying keys to external storage addresses include: hashing, the mathematical manipulation of key numbers often including division by a prime number; folding, the adding of first and last halfs of key numbers; splitting, the use of some subset of a key number; and coding, the replacement of non-numeric with numeric characters to form an address.

Support for this dichotomy of access mechanisms can be seen in the implementation of I.S.& R. systems. Automation of one does not require that the other be automated as well. In fact, the most common automated I.S.& R. system design produces a list of relevant documents and the locations from which they must be manually retrieved.

In some systems the intellectual access step generates not only the names and locations of relevant information items, but also summaries, extracts or abstracts describing the items determined to be pertinent by the indexing mechanism. Examples of this occur in the manual case when an annotated reference such as a subject area abstract is used as well as in automated systems. In the former case, the additional information can be viewed as being generated by the intellectual access process. In the automated case, however, the retrieval and production of abstracted information is likely to be an additional physical access step, retrieving summary information from a file of abstracts. In either case, the user can selectively determine the information items which he wishes to physically access.

For evaluating user satisfaction, the entire retrieval process must be considered; both intellectual and physical access are required to satisfy user needs. Consequently, any I.S. & R. performance measure which does not take into account both types of access can not be considered a good reflection of user satisfaction. We consider search length as a measure of performance and a surrogate for user satisfaction, first for the intellectual access and then for the physical access component of an I.S. & R. system.

Intellectual Access Model

The criteria of search length, the number of documents that must be scanned before a predetermined number of relevant documents is found, has been proposed as a surrogate for user satisfaction. Cooper (8) presents assumptions which, although not totally valid, lead to the acceptance of search length as an approximate measure of I. R. & S. system performance:

1. For any specific user, one needs a fixed number of "relevant" documents and will continue to search until all of these have been found or until the collection has been exhausted.

2. A document is relevant if it has some usefulness to the user.

3. The average utility of a retrieved document is equal for all retrieval systems and is:
    a. positive if the user deems the document useful
    b. negative if the user deems the document not relevant.

4. The utility of an unretrieved document is zero because the user never sees it nor is he concerned with its relevance.
5. The utility of a retrieved document is independent of the order of the search sequence, the order in which the documents are presented to the user.

Using these assumptions, we can see the efficacy of using search length to operationalize the criterion of user satisfaction as applied to the intellectual access component of an I.S. & R. system.

Let $U_{ij}^{(r)}$ = utility to user i of retrieval system j as a function of r,

Where: r = number of relevant documents desired by the user (we assume there are at least r relevant items in the collection).

Cooper shows that: (we present our own notation for clarity)

$$U = cr + d(n-r)$$

Where: c = average utility of a relevant document retrieved

d = average utility of an irrelevant document retrieved

n = number of documents searched

We have assumed that c is positive, d is negative, and that $r \leq n$. This implies that $U' = U/d = (c/d)r + (n-r) = c'r + n$, where $c' = c/d - 1$. If r, c, and d are fixed and known, as is assumed, then maximization of U implies minimization of U', which in turn implies a minimization of n, the search length. Since r is fixed, this also implies a minimization of n - r, Cooper's idiosyncratic definition of search length which is in reality the number of nonrelevant documents a user must wade through

to get r relevant documents. Note that the problems of estimation of the parameters of c and d are alleviated, and that Cooper's search length is really independent of r, with n the variable of interest. Finally, we recognize that search length is also intuitively related to response time and cost.

In a second paper, Cooper (6) develops what he calls a design equation which relates search length to various IS&R system parameters. He distinguishes this from a performance equation which defines a measure of effectiveness and can be used to predict performance for a specified set of parameters. The design equation cannot be used for prediction as it involves variables whose values are unique to a given retrieval situation and cannot be predicted over time. Moreover, these variables are stochastic, and their distributions and moments are complex and interdependent; they are therefore unknown a priori to the analyst. Cooper suggests using the design equation to calculate the expected search length of a given retrieval system configuration for each of a sample of user queries and calculate the average search length for the system. This average value can then be used in comparing the utilities of alternative system configurations. His model assumes that there is one and only one relevant document in the collection and that this is the only one desired for a specific user query (r = 1).

Let: M = the number of documents in the collection,

L = the number of terms in the limited, fixed indexing language vocabulary,

N = the number of specific user queries under consideration,

a (i, ℓ) = a binary variable representing the document indexing function, being equal to +1 if the ℓth descriptor is assigned to document i, and equal to -1 if not,

$r(j)$ = the subscript referring to the one document in the collection that is relevant and desired for query j,

$q(\ell,j)$ = a tertiary variable representing the query analysis function, being equal to +1 if descriptor ℓ is used in query j in a positive sense, being equal to 0 if descriptor ℓ is not used in query j at all, and being equal to -1 if descriptor ℓ is used in query j in a negative sense (with a 'not' connector so descriptor ℓ becomes a negated term), and

$J(j)$ = set of descriptors assigned to query j = $\{\ell \mid |q(\ell,j)| \geq 1\}$.

We see that (again modifying Cooper's notation):

$$S(j) = \|J(j)\| = \sum_{\ell=1}^{L} |q(\ell,j)|$$

= size of query j in terms of the number of descriptors assigned to it:

Where: $|\alpha|$ = absolute value of $\alpha = \begin{cases} \alpha \text{ for } \alpha \geq 0 \\ -\alpha \text{ for } \alpha < 0 \end{cases}$

$\|\beta\|$ = number of items in set β.

Cooper then defines and formulates the variables:

$$I(K) = 1/M \sum_{i=1}^{m} \prod_{\ell \in k} a(i,\ell)$$

= imbalance among a specified subset, k, of the L possible descriptors with respect to a given document collection;

$$E(j) = \tfrac{1}{2} \sum_{\ell=1}^{L} \left[|q(\ell,j)| - q(\ell,j)\, a(r(j),\ell) \right]$$

= error of query j when matched against the collection in terms of the specified number of mismatches or descriptors that do not match between the query and the one relevant document, $r(j)$, in the collection for that query;

$$O(j) = S(j) - E(j)$$

= overlap for query j in terms of the number of descriptors that do match between the query and the one relevant document for that query.

Finally, Cooper demonstrates that the expected search length for query j is determined by:

$$(1)\quad ESL(j) = \tfrac{1}{2} + M \sum_{k \le S(j)} H(k,j)\, I(k) \prod_{\ell \in K} q(\ell, j)$$

$$\text{Where: } H(K,j) = 2^{(-S(j)-1)} \sum_{z=0}^{E(j)} \binom{\|k\| - 1}{z} \binom{S(j) - \|k\| + 1}{E(j) - z} (-1)^z ,$$

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} , \text{ and}$$

$$a! = a\,(a-1)(a-2)\ldots 3 \cdot 2 \cdot 1 .$$

The average search length for a given system is then given by:

$$\overline{ESL} = 1/N \sum_{j=1} ESL(j)$$

Some explanation is required. Consider the set of M documents in the collection as unordered. The matching of the query to the document index orders the documents intellectually, into a search order of S(j) + 1 levels. This is, as Cooper states, the weak ordering action upon the collection by the retrieval function. Each level corresponds to the documents that have an equal number of mismatches or errors; with e representing the number of errors on a particular level. If e = 0, one is at the highest level with all documents here having a perfect match for all S(j) descriptors in query j; if e = 1, one is at the next level down with all documents here having one mismatch and S(j)-1 matches; and so on until we reach e=S(j) where all documents on this level have no matches with the query. Assuming the search is based on reasonable faith in the indexing scheme, one starts at level e = 0 and randomly searches among the documents on this level until the one relevant one is found or the level is exhausted. One then goes to the next level down and searches until the document desired is found. We note that e = E(j) is not always zero since no retrieval system is always perfect. The number of

documents on level e is given by Cooper as

$$\sum_{\substack{y \in J(j) \\ |y| = e}} \sum_{i=1}^{M} \prod_{\ell \in J(j)-y} \tfrac{1}{2}\left[1 + q(\ell,j)\right] a(i,\ell)\Big] \prod_{\ell \in y} \tfrac{1}{2}\left[1 - q(\ell,j)\, a(i,\ell)\right]$$ where

y is a subset of size e of the set J (j) and J (j)-y is the subset of terms in set J (j) that are not in subset y. Cooper counts the number of documents in the levels e = 0, e - 1,..., e = E (j)-1 and then adds to that the expected number of documents searched on level e = E(j), the level where the relevant document is. A linear search is carried out on the documents of level e = E(j) where one assumes the relevant document is mixed in at random with the other documents on this level. This type of search will be discussed below when we look at physical access.

Some final points are in order. Cooper's expected search length is not the same as ours; it is one unit smaller since he does not count the relevant document in his search length and we do. Another point of note is that the a (i, ℓ) variables, specific to the document collection and indexing language, the q(ℓ,j) variables, specific to the user query set and the indexing language, and the E(j) variable, specific to the collection and the query set and the indexing language, are all part of the search length design equation and cannot be predicted for a general situation. Thus, one cannot use the design equation for prediction, but only for explanation. Finally, minimization of search length will imply a minimization of E(j), in terms of optimization of the indexing language, with all other things held constant.

This section primarily consists of a survey of Cooper's work on applying search length as a measure of IS&R system performance. We have recognized, however, that his frame work and design equations most closely describe the first component of such a system, intellectual access; thus, we view the variables C, d, n, r, and ESL(j) in terms of information items "retrieved" in the sense that they appear on a list of pertinent references generated

by an intellectual access module. In the next section we consider search length as a measure of performance for the physical access component of an IS&R system.

Physical Access Models

Consider a linear array of the M documents in the library storage area. A user has gained intellectual access to the documents in terms of knowing of their existence and of being able to recognize the relevant ones as those he wants upon some sort of perusal. Let us consider a series of variations in the structure or organization of the storage facility and compare the effects on expected search length.

The first model assumes that all M documents have been placed on shelves or whatever in storage in random order. Moreover, no prior knowledge of the location is known. This is what Chapin (4) calls a strict sequential file. Let R = the number of relevant documents in the collection that could be useful, and recall that r relevant documents are required and n documents are retrieved. Of course, we must have $M \geq R \geq r$. This yields: $Pr(n) = \left[\binom{R}{r-1}\binom{M-R}{n-r} / \binom{M}{n-1}\right] \times (R-r+1)/(M-n+1) = \left[\binom{R}{r}\binom{M-r}{n-r} / \binom{M}{n}\right] \times \frac{r}{n}$ = the probability of a search length of size n. This means that

(2) $EXP(n) = \sum_{n=r}^{N} nPr(n)$ = expected search length = $r(M+1)/(R+1)$, as Cooper (6) proves. We note the special case where R=r=1 gives rise to the familiar search length for the one and only one document of $(M + 1)/2$. This is the structure Cooper uses in his design equation formulation to calculate the search effort on level e=E(j).

Suppose one wishes to reduce search length by grouping documents together into P clusters. Each cluster has documents deemed relevant to a specific concept or discipline or field of knowledge. Note that we are here concerned with the clustering of the documents themselves in storage rather than with

the indexing mechanism discussed in the previous system. This clustering does correspond to the weak ordering action of the retrieval function.

Note that the idea of grouping to save on search length was used by Leimkuhler (13). He showed that if one can assign probabilities of relevance for a given query to each document, the best sequence in terms of minimum search length is to search through those documents that are most likely to be relevant first; i.e., the search sequence is in the order of declining probabilities of relevance. Baker (2) considers the search sequence if one wants to minimize the total cost of the search time. He shows that, for one specific query, one searches documents in the rank order based on the decreasing ratio of the probability of relevance of a specific document to the cost of searching that specific document. This also leads to reward maximization if the reward is independent of the relevancy of the document. He further considers the proper ordering of documents when the collection will be confronted with a given set of queries, and the effects of a constraint on the total time available to do a search. Finally, he considers a parametric analysis to see how competition and service modifications change the search length.

Morse (14) considers searching a collection arranged in some order in order to seek a set of desired documents by browsing. Thus, there is no intellectual access before the physical search begins. He also considers grouping and searching the most likely groups first to minimize the number of relevant documents found when the user has only a limited amount of time to spend on browsing. Morse looks at a browsing search rate, based on models from submarine search theory, as a function of glimpse rate and glimpse accuracy. He also considers the efficacy of linking descriptors to form a query profile.

Leimkuhler points out that since the measurement of such probabilities for a strong ordering may be impossible, the assignment of probabilities to groups in a weak ordering, such as Cooper (5) considers and uses, may be useful. Here one searches through the groups in rank order of the decreasing probabilities of relevance, searching randomly through the documents on each level. Leimkuhler considers P, the number of groups into which one divides the collection, and shows the effect of that number on search length for given probability distributions, although he never shows that the "best" value of P would be.

We assume that the grouping recall is perfect in that all relevant documents are in the same group. Without this assumption one is left with an indexing error rate and the design equation format that Cooper (6) used. This means that R= the size of the group that has the relevant documents. The average group size is M/P. Thus, we shall assume that R= M/P. The searcher scans the group headings, with the groups arranged in random order, until the desired group is found and then a random search among the documents occurs. The expected search length is given by the following equation. (3) $EXP(n) = (P + 1)/2 + (M/P + 1)r/(r + 1)$. The first term in equation (3) corresponds to the expected number of group headings searched, and the second term corresponds to the expected number of documents searched within the relevant group. In this and all following cases of physical access, we shall add the group heading and document searches together to get one grand total for search length. If there are no group headings, a case we do not consider in this paper, so that one must scan a document in each group to determine which group one has, and if one assumes that such a scan will reveal if the group is relevant, one must subtract one from EXP(n).

Suppose, to further reduce search length, one orders the groups by some sorting technique. This is not an ordering by likelihood as Leimkuhler, Baker, and Morse discuss, but a sorting in some sequence to facilitate search. An example is the yellow pages of the telephone directory where the listings are grouped by service and/or product offered, and the group headings are alphabetized. We shall assume that the group heading desired is known, via intellectual access. This enables us to use a binary search procedure, outlined below, which minimizes the search length of an ordered linear sequence if one assumes that each group has an equally likely chance of being the desired group.

The binary procedure starts by looking at the $\lfloor (P + 1)/2 \rfloor^{st}$ group heading, where $\lfloor \alpha \rfloor$ = the largest integer $\leq \alpha$. If this group, in the middle of the set, is the desired group, stop. If not, it is either in the subset of groups above the one just rejected or the subset below; due to the ordering, the relevant group subset is immediately known. The routine is then repeated upon the group subset containing the desired group.

We note that this procedure cuts off half of the remaining groups in each iteration, hence the name "binary" search. The expected number of group headings to be searched is given by Knuth (11):

$$(1+1/p)(\lfloor \log_2 P \rfloor + 2) - \left(2^{\lfloor \log_2 P \rfloor + 1}/P\right) - 1.$$

This means, if we assume that the documents within the group are randomly distributed, unlike our yellow pages example, that

$$(4)\quad EXP(n) = \left[(1+1/P)(\lfloor \log_2 P \rfloor + 2) - 2^{\lfloor \log_2 P \rfloor + 1}/P - 1\right] + \left[(M/P + 1)\, r/(R + 1)\right].$$

The first bracketed term in formula (4) gives the expected length for the binary search of ordered groups and the second term yields the expected sequential search length among the unordered documents in that group.

Finally, consider the case where not only are the groups ordered but the documents are also ordered within each group. The yellow pages are such an example. If one relevant document exists and only one is desired, R = r = 1, the expected search length is given by

$$(5)\quad EXP(n) = \left[(1 + {}^{1}/_{P})(\lfloor \log_2 P \rfloor + 2) - 2^{\lfloor \log 2P \rfloor + 1}/P - 1\right] + \left[(1 + {}^{P}/_{M}) \cdot (\lfloor \log_2 ({}^{M}/_{P}) \rfloor - 2) - P\left(2^{\lfloor \log_2 M/P \rfloor + 1}/M\right) - 1\right].$$

Again, the first term in brackets corresponds to the group heading search and the second term to the search among documents in the desired group. In equation (5) however, the binary search algorithm is used to locate both the desired group and the document within the group. The situation where R and/or r is greater than one has not been modeled yet, but poses an interesting problem in a modified binary search procedure of some type. It would also be interesting to consider the effects of searching when it is not assumed that all documents in a group are equally likely to be relevant.

We now present an example to demonstrate these formulas. Consider a collection of M = 1024 documents that may be grouped into P = 64 groups and three sets of relevance parameters: a) R = 1, r = 1;

b) R = 16, r = 1;

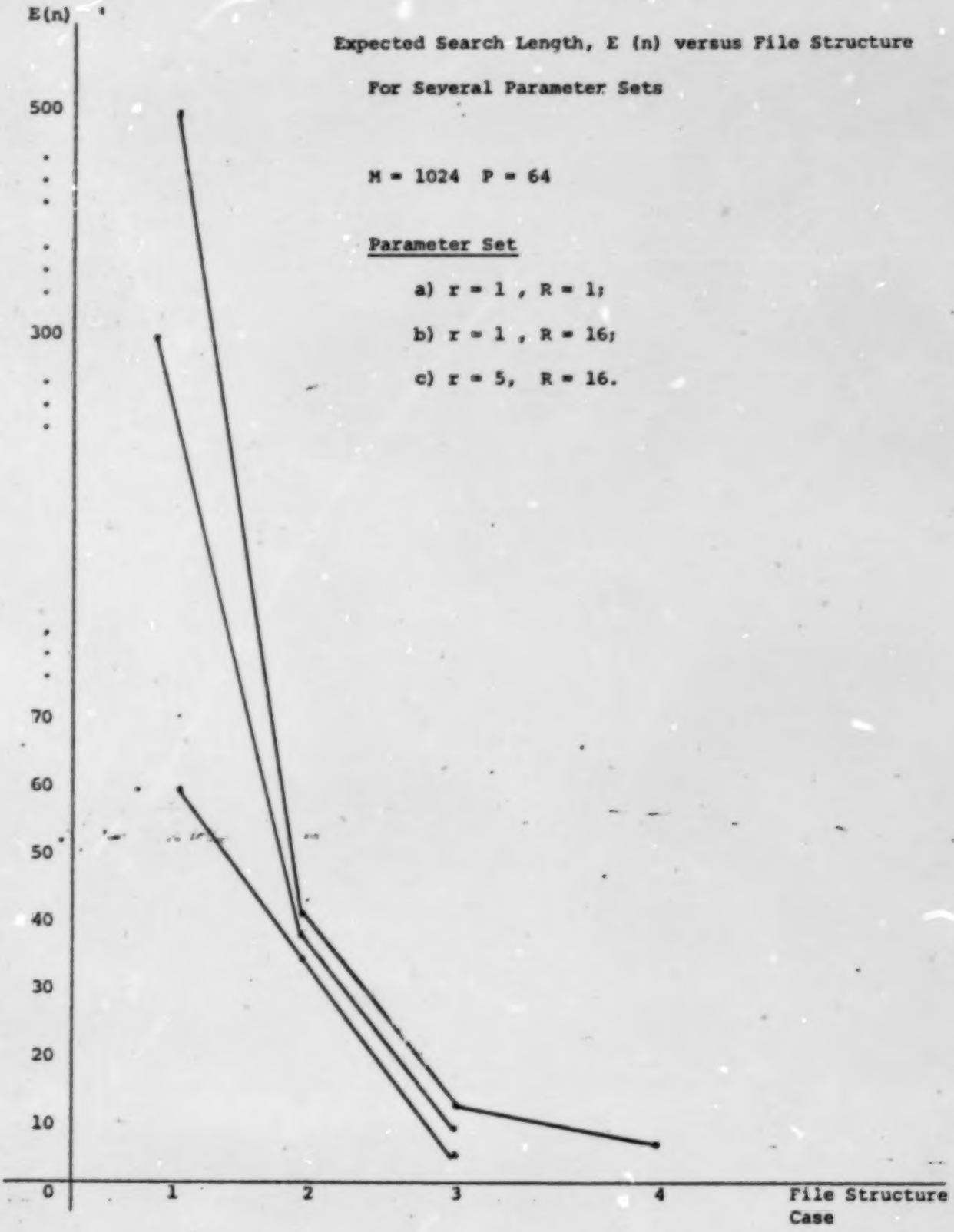
c) R = 16, r = 5.

The physical search length derived from the appropriate formula for each file structure case is given in Table 1, and is illustrated in Figure 2. Note how increased indexing effort, in terms of the grouping and sorting, reduces the search effort, in terms of the expected search length, for all three parameter sets. Moreover, comparing parameter sets a) and b) shows that an increase in the number of relevant items in the collection, R, decreases expected search length for all file structure cases. Finally, comparison of parameter sets b) and c) indicates that an increase in the number of relevant

documents desired, r, results in an increase in the expected search length.

REF	FILE STRUCTURE	FORMULA	PARAMETER SET a) R=r=1	b) R=16, r=1	c) R=16, r=5
1	Random	2	512.5	60.3	301.5
2	Grouped	3	41.0	33.5	37.5
3	Ordered Groups	4	13.625	6.125	10.125
4	Ordered Groups, Ordered Within	5	8.50	-	-

Table 1 - Expected Search Length, E (n) with M = 1024, P = 64



Expected Search Length, E (n) versus File Structure

For Several Parameter Sets

M = 1024 P = 64

Parameter Set

a) r = 1 , R = 1;

b) r = 1 , R = 16;

c) r = 5, R = 16.

Figure 2

Towards an Overall Design Model

This paper examines search length as it applies to the intellectual and physical access components of an information storage and retrieval system. The whole raison d'etre for considering search length is that it serves as a surrogate measure of user satisfaction, and indirectly of system performance. Ultimately, an overall design model is a desirable goal. Our current knowledge of the complex inter-relationships between IS&R system variables, unfortunately, only allows us to speculate about the general form such a model will take.

Search length reflects only the effort required to identify and retrieve relevant documents. We have seen that as indexing improves, search length related to intellectual access decreases. In addition, when information items in storage are grouped and ordered physical access search effort declines. Because indexing and the structuring of information in storage are not independent and also represent effort, we cannot assume that an optimal system will result from an attempt to minimize search length by building more complex and costly indexing retrieval systems. Other types of effort are also implicit in this IS&R system conceptualization. For instance, update effort is required to add new information to the collection, to correct indexing for items already in the system, and to delete information items.

One possible model formulation would minimize total search effort as a function of expected search length, subject to constraints on the amount of time, money, space, and other scarce resources which can be applied to the various activities such as indexing, storing, updating, and the like.

The model here would be:

$$\text{MIN } E(s) = F_s(ESL_I, ESL_P)$$

$$\text{S.T. } \sum_j f_j(E_{ij}) \leq r_i \; \forall i$$

Where E (s) = total search effort

ESL_I = intellectual access expected search by the

ESL_p = physical access search length

F_s = the total search effort function

E = the effort for activity j in terms of resource.

r = the available units of resource

An alternative formulation would minimize the total retrieval effort, subject to constraints specifying maximum acceptable levels of performance in terms of search length:

$$\text{MIN } TE = F_T(E;i)$$
$$\text{S.T. } E(s) \leq MS$$

Where: TE = total effort

MS = the maximum acceptable search effort

F_t = the total retrieval effort function

The complexity of this objective function is greater than in the previous formulation because it considers the effort of different individuals and components of the library system. For example, search effort is increased mostly by the user while indexing and update are performed by the personnel and/or machinery of the retrieval system.

Although models like those outlined above will be useful, it is likely that design models will have to consider various performance criteria in addition to search lengths as a representation of user satisfaction. The incorporation of multiple criteria into future models may be achieved in various ways including the use of expanded constraint sets or by application of modeling techniques such as multiple goal programming.

In all of the formulations above, the important decision variables deal

with the amount of effort applied to indexing and file structuring. Implementation of models of this type will require an understanding of the effects of these variables on IS&R system performance as well as a methodology for their measurement. Because of the stochastic nature of the variables, their complexities and interdependencies, any future overall design optimization models will probably be parametric.

Prior to the successful development of an overall model for improving IS&R system design, we can apply current knowledge to improve system performance. Changes designed to reduce perusal time rather than the number of perusals in a library situation can be productive (1). Shelf markers should be easy to see and read. Document search is facilitated by placing labels on the spines of books and arranging the books properly on the shelf; the user can then scan, browse, and peruse the keys quickly and easily. Good lighting also helps. The application of the relationships uncovered in the process of developing quantitative models is also possible. The user must be mindful, however, of the current incomplete understanding of variable interrelationships. We have demonstrated, for instance, that grouping and ordering of information items reduces physical access search length; the effects of this strategy on indexing, updating and other effort, on performance measured by criteria other than search length, or on resources, are not known at this time. Simulation, first of simplified conceptualization such as the one presented in this paper and then of models with increasingly more criteria and variables, promises to produce early answers to design questions.

BIBLIOGRAPHY

1. Bates, M., Personal Communication, University of Maryland College of Library and Information Services, College Park, Maryland, 1973.

2. Baker, N. R., "Optimate User Search Sequences and Implications for Information Systems Operation", technical paper, Purdue University School of Industrial Engineering, Lafayette, Indiana, 1968.

3. Becker, J. and Hayes, R., Information Storage and Retrieval: Theories, Models and Tools, Wiley, New York, 1968.

4. Chapin, N., "Common File Organization Techniques Compared", Proceedings, Fall Joint Computer Conference, 1969.

5. Cooper, W. S., "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems", American Documentation, vol. 19, 1968, pp. 30-41.

6. Cooper, W. S., "On Deriving Design Equations for Information Retrieval Systems", Journal of the American Society for Information Science, vol. 21, 1970, pp. 385-95.

7. Cooper, W. S., "A Definition of Relevance for Information Retrieval, Information Storage and Retrieval, vol. 7, 1971, pp. 19-37.

8. Cooper, W. S., "On Selecting a Measure for Retrieval Effectiveness, Part I. The 'Subjective' Philosophy of Evaluation, Part II. Implementation of the Philosophy", Journal of the American Society for Information Science, vol. 24, 1973, pp. 87-100, 413-24.

9. Evans, E., H. Borko, P. Ferguson, "Review of Criteria Used to Measure Library Effectiveness", Bulletin of the Medical Library Association, vol. 60, 1972, pp. 102-10.

10. Greenberg, H. J., Personal Communication, Management Science Systems, Rockville, Maryland, 1973.

11. Knuth, D., The Set of Computer Programming, Vol. 2, chap. 1 & 2; vol. 3 chap. 5 & 6, Reading, Mass., Addison-Wesley, 1973.

12. Lancaster, F.W., "The Cost-Effectiveness Analysis of Information Retrieval and Dissemination Systems", Journal of American Society for Information Science, Vol. 22, 1971, pp. 12-27.

13. Leimkuhler, F.F., "A Literature Search and File Organization Model", American Documentation, vol. 19, 1968, pp. 131-6.

14. Morse, P.M. "Search Theory and Browsing", Library Quarterly, vol. 40, 1971, pp. 391-408.

15. Neuts, M., Personal Communication, Purdue University Department of Statistics, Lafayette, Indiana, 1973.

16. Orr, R.H., "Progression Documentation: Measuring the Goodness of Library Services: A General Framework for Considering Quantitative Measures", Journal of Documentation, vol. 29, 1973, pp. 315-332.

17. Papier, Lawrence, "Sources of Criteria for Intermediate Evaluation of Information Systems", Management Informatics, Vol. 2 (1973), No. 2, pp. 71-74.

18. Shaffner, R., Personal Communication, Richard Abel and Co., Portland, Oregon, 1973.

19. Travis, I., Personal Communication, University of Maryland College of Library and Information Services, College Park, Maryland, 1973.